Distant Reading in Digital Humanities:
Case Study on the Serbian Part of the ELTeC Collection

Ranka Stanković*,
Cvetana Krstev†, Branislava Šandrih Todorović†,
Duško Vitas‡, Mihailo Škorić*, Milica Ikonić Nešić†

*University of Belgrade, Faculty of Mining and Geology, Serbia
{ranka, mihailo.skoric}@rgf.bg.ac.rs
†University of Belgrade, Faculty of Philology, Serbia
cvetana@matf.bg.ac.rs, {branislava.sandrih, milica.ikonic.nesic}@fil.bg.ac.rs
‡University of Belgrade, Faculty of Mathematics, Serbia
vitas@matf.bg.ac.rs

LREC, June 2022

Outline

Distant Reading COST Action (CA16204)

- Distant Reading for European Literary History (2017–2022);
- Creates a network of experts whose aim is to develop resources and tools that can help in writing the European literary history;
- Main objective is the production of an unified, uniform, multilingual, digital novel collection dubbed ELTeC *European Literary Text. Collection*

Position of the Serbian language

- Serbian part of ELTeC, dubbed SrpELTeC sub-collection;
- In this paper, we explained all the work that has been done so far, including:

Position of the Serbian language

- Serbian part of ELTeC, dubbed SrpELTeC sub-collection;
- In this paper, we explained all the work that has been done so far, including:
  - novel selection, text preparation, structural annotation;

Position of the Serbian language

- Serbian part of ELTeC, dubbed SrpELTeC sub-collection;
- In this paper, we explained all the work that has been done so far, including:
  - novel selection, text preparation, structural annotation;
  - POS-tagging, lemmatization and named entity recognition;

Position of the Serbian language

- Serbian part of ELTeC, dubbed SrpELTeC sub-collection;
- In this paper, we explained all the work that has been done so far, including:
  - novel selection, text preparation, structural annotation;
  - POS-tagging, lemmatization and named entity recognition;
  - visibility and accessibility of the SrpELTeC;

Position of the Serbian language

- Serbian part of ELTeC, dubbed SrpELTeC sub-collection;
- In this paper, we explained all the work that has been done so far, including:
  - novel selection, text preparation, structural annotation;
  - POS-tagging, lemmatization and named entity recognition;
  - visibility and accessibility of the SrpELTeC;
  - SrpELTeC as a resource for research in DH.

Towards the new collection

- Each language sub-collection should contain novels originally written in that language and first published in the period 1840–1920.
- For this purpose, a "novel" is defined as a fictional narrative text at least 10,000 words long.
- The choice of novels cannot be random, since some balancing criteria have also to be met.

Criteria for the new collection

- Should contain 100 works that qualify as ‘novels’;
- Should as-much-as-possible equally represent both male and female authors;
- Should cover the whole time period uniformly;
- Should contain novels of different sizes: short ($\leq$ 50,000 words), medium ($\geq$ 50,000 words and $\leq$ 100,000) and long ($\geq$ 100,000 words);
- Should contain both well-known canonical works and less-known and forgotten novels;
- Should represent optimally 9–11 authors by three novels, while all other authors should be represented by one novel only.

What about Serbian?

Clearly, the development of language sub-collections would not be the equally demanding task for all languages.

Challenging conditions

First, for some languages most of the novels from the chosen time period were not yet digitized or their digitized versions were not available.
Second, novels as a literary form were just emerging in the chosen time period, leading to the difficulty to fulfill the demanding balance criteria.

Novel selection and text preparation

- The first task was to create a list of candidate novels (consulted Mutual library catalog of the Republic of Serbia);
- Ending up with a list of 157 candidate literary works, digitized versions were looked up;
- It turned out that mostly canonical works were digitized but not in a proper manner, i.e. without proper metadata, so we had to compile the whole sub-collection from scratch (retrieve hard copies, scan them and do the OCR, proofread and correct them and do the basic annotations, the so-called level-1 annotation);
- Eventually, 100 candidates were chosen for the SrpELTeC sub-collection, and remaining ones were put in the extended sub-collection SrpELTeC-ext.

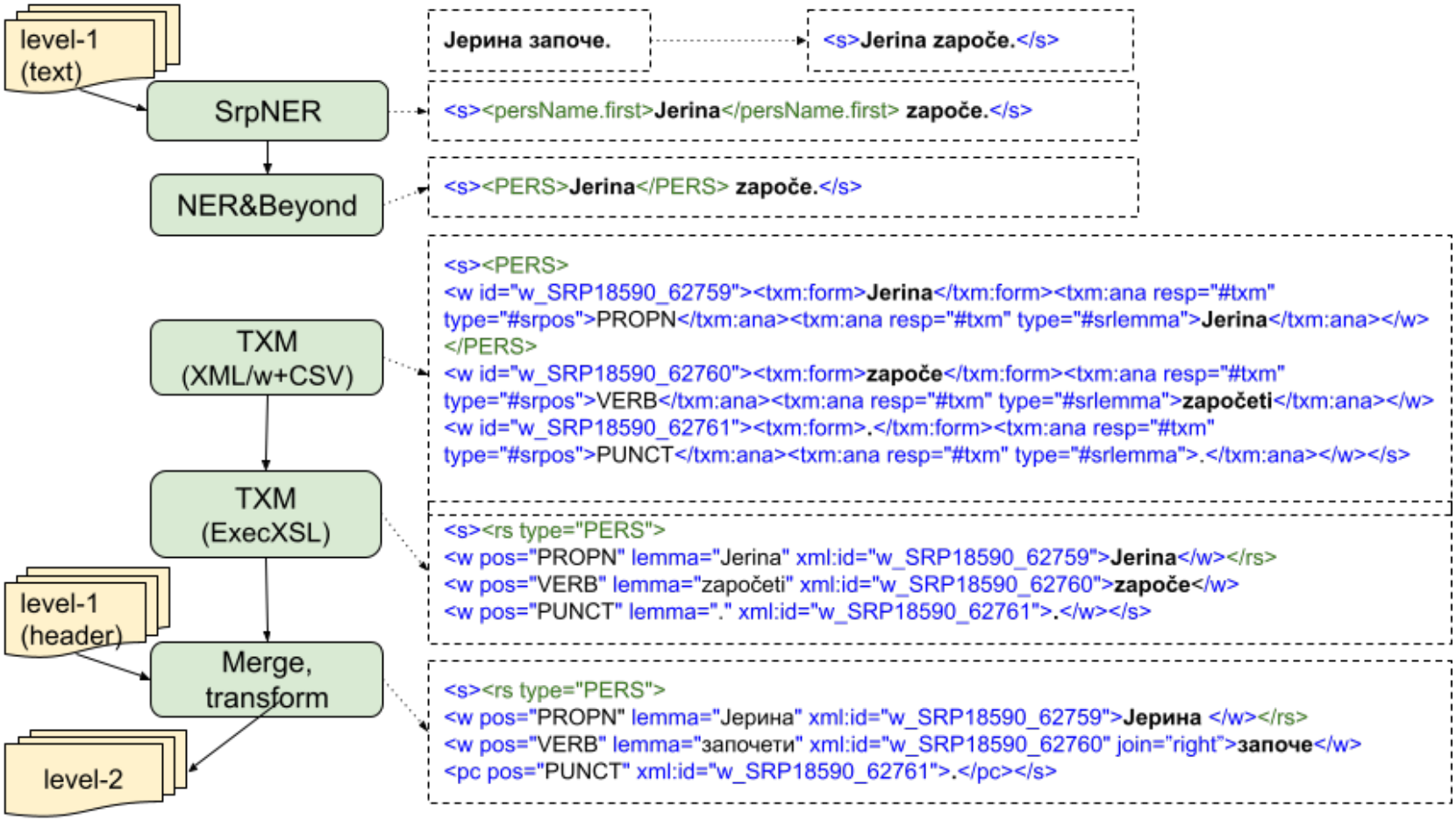
Novels as XML/TEI documents

- Each novel from the ELTeC collection at level-1 was prepared as an XML/TEI document;
- Having consistent sub-collection headers enabled various further analysis: titling practices of Serbian narrative literature, authors' gender and age, publication places, modes of publication, etc.

More complex level-2

- The level-2, built on the basis of level-1, is more complex and informative (sentence tags, token tags for words and for punctuation separately, attributes for word's part-of-speech and lemma, etc.);
- Annotation pipeline built upon various language resources and tools was designed and developed.

Annotation pipeline



NER, POS-tagging and lemmatization

- The pipeline starts with NER, for which the rule- and lexicon-based SrpNER system was used that recognizes various classes;
- Since the level-2 tagset should contain PERS, ROLE, LOC, ORG, DEMO, EVENT and WORK tagset annotations, we had to adjust our SrpNER tagset;
- TXM is using an appropriate parameter file for TreeTagger, used for the part-of-speech tagging and lemmatization; therefore, TreeTagger's lexicon also had to be prepared.

SrpELTeC-gold & SrpCNNER

- First outcome of this process is the named-entity (NE) annotated corpus SrpELTeC-gold, publicly available on the European Language Grid (ELG) platform;
- The second one is the NE Recognizer, dubbed SrpCNNER, that was trained using spaCy (Python module for advanced NLP) on SrpELTeC-gold to recognize seven previously mentioned NE types with a Convolutional Neural Network (CNN) architecture; The model achieved F_1 score of $\approx 91\%$ on the test dataset.

Digital Libraries and Corpus Querying

SrpELTeC is publicly available on three platforms:

Udaljeno čitanje Original pages as pictures while reading in parallel a digitized version;

Aurora Gives detailed insight into the vocabulary of novels, offering them to browse texts, concordances and frequency lists;

Sketch Engine Well-known platform for corpora management and exploration.

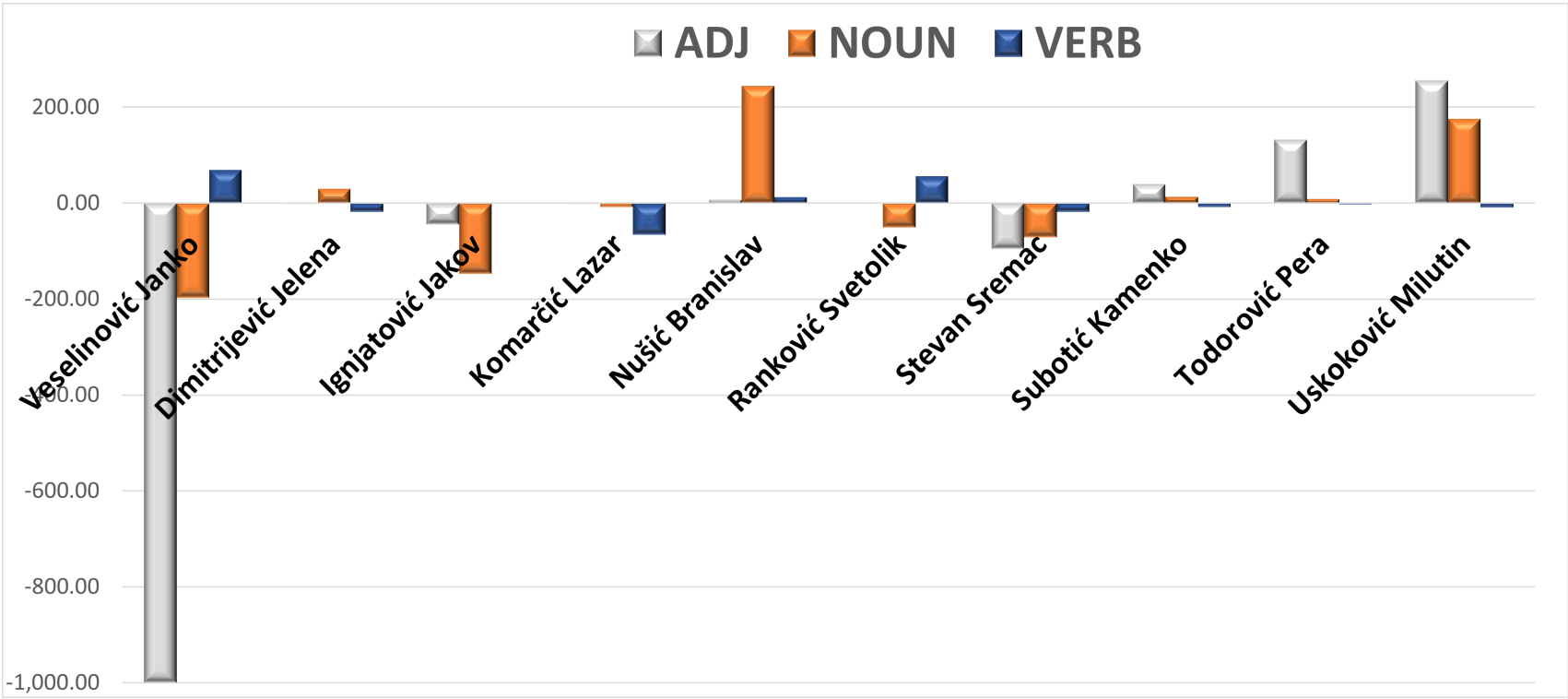
SrpELTeC in Wikidata

- Novels preparation for Wikidata and linking Wikidata to various applications has first been done manually;
- The automation procedure followed using OpenRefine and QuickStatements;
- Information about authors, novel titles, publication places and other metadata with different visualization options were implemented on top of SrpELTeC Wikidata, and can be retrieved by a corresponding SPARQL query.

Textometric analysis

- TXM tool contains modules that calculate various text statistics:
  - KWIC concordances of word patterns based on CQP full text search engine and CQL query language;
  - word pattern frequency lists based on tokens, lemmas, POS, or structural annotations including NEs;
  - word pattern progression graphics;
  - rich HTML-based text edition navigation with links from all other software modules.

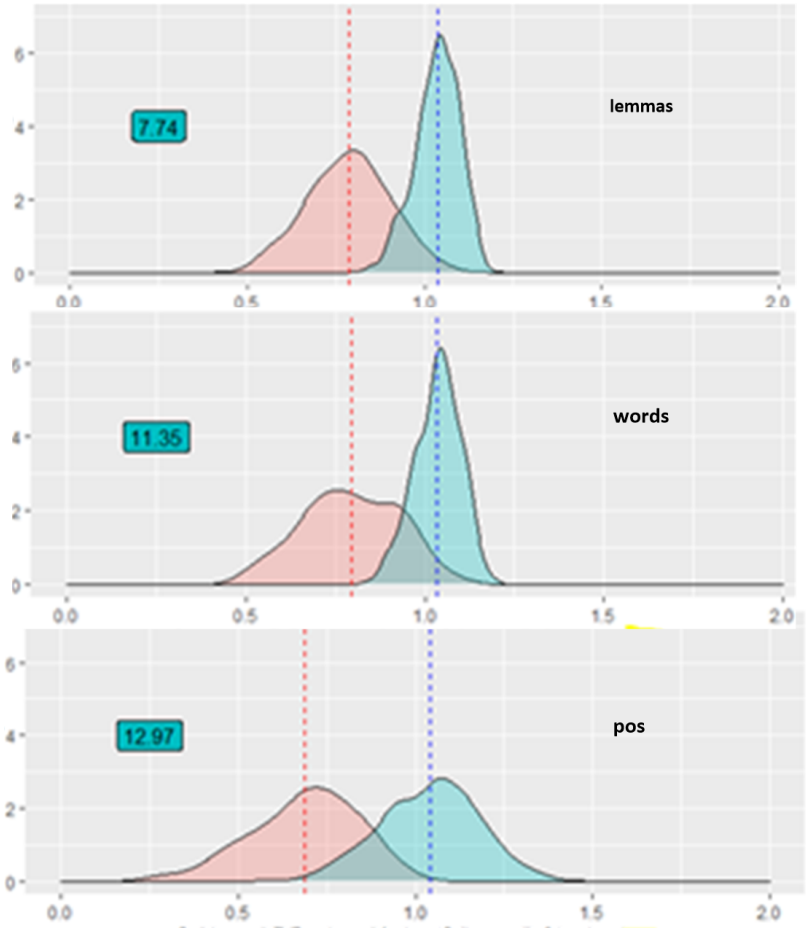
Findings about the SrpELTeC



SrpELTeC application in digital humanities

- Research on comparative stylistic and morphosyntactic analysis of ELTeC texts followed;
- Goal was to find out what better defines an author's style: the use of word forms, lemmas, or certain part of speech;
- Based on metadata, analysis of the influence of author's sex or a time period was also investigated.

Authorship attribution



Similes in SrpELTeC

- SrpELTeC served as an inspirational foundation for analysis of the use of rhetorical figures in Serbian literary texts;
- Such is the presence of simile figures, due to their prevalence and recognizable structure;
- To retrieve (and annotate) simile figures from any Serbian text, we developed a local grammar in the form of the expandable FSA that relies on the Serbian morphological e-dictionaries.
- The most frequently used similes are
  - *beo kao sneg* (white as snow) (used in 30 novels);
  - *bled kao krpa* (pale as a cloth);
  - *bled kao smrt* (pale as death);
  - *hladan kao led* (cold as ice).

SrpELTeC as the evidence of eating habits

- SrpELTeC was a valuable source for studying everyday life in the second half of the 19^{th} and the beginning of the 20^{th} century, and especially their eating habits;
- Some of the questions posed were:
  - what were the most popular foodstuff and meals;
  - how they were named;
  - and how these changed over time.
- For this research the Serbian morphological e-dictionaries implemented in Unitex were indispensable since various foodstuff, meals and drinks.

- In this paper we presented our share in the project, namely SrpELTeC sub-collection;
- We have shown that although it was primarily developed to be used for distant reading methods and tools, it was equally useful for close reading approaches;
- Our next activities are:
  - Automation of NEL to Wikidata by training a suitable machine learning model;
  - Expansion of the NE tagset with other entities, like drinks, or transportation means;
  - Assignment of semantic attributes present in electronic dictionaries to the Aurora's concordances;
  - Finally, the annotated level–2 corpus will be published in the Linguistic Linked Open Data.

Thank you for your attention!

Questions?